

DOCUMENT RESUME

ED 264 262

TM 850 746

AUTHOR Levine, Michael V.
TITLE The Trait in Latent Trait Theory.
SPONS AGENCY Office of Naval Research, Washington, D.C.
PUB DATE Jul 82
CONTRACT N00014-79C-0752; NR-150-445
NOTE 26p.; In: Item Response Theory and Computerized Adaptive Testing Conference Proceedings (Wayzata, MN, July 27-30, 1982) (TM 850 744).
PUB TYPE Reports - Research/Technical (143) --
Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Item Analysis; *Latent Trait Theory; Mathematical Models; *Measurement Techniques; Psychological Testing; *Psychometrics; Scoring; Statistical Analysis; *Test Items; *Test Theory

ABSTRACT

Significant to a latent trait or item response theory analysis of a mental test is the determination of exactly what is being quantified. The following are practical problems to be considered in the formulation of a good theory: (1) deciding whether two tests measure the same trait or traits; (2) analyzing the relative contributions of a pair of traits or abilities to test performance; (3) detecting "functional" changes in items including those caused by security problems, modes of administration changes, and changes in familiarity with the concepts supporting the item in the population being tested; (4) determining the adequacy of an item response function, that is, a specific mathematical formula relating performance to ability; (5) discovering the shape of the item response functions, including multidimensional item response functions; (6) quantifying the magnitude and reliability of violations of the principal assumption of latent trait theory, "local independence"; and (7) modeling item responses (such as omitting or changing answers) that fail to be locally independent. This paper outlines theoretical results bearing on these seven problems. A new theory is presented, with the central problem the representation of traits, abilities or achievements, and their distributions. (PN)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

THE TRAIT IN LATENT TRAIT THEORY

MICHAEL V. LEVINE
UNIVERSITY OF ILLINOIS

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☒ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Significant to a latent trait or item response theory analysis of a mental test is the determination of exactly what is being quantified. The following are practical problems to be considered in the formulation of good theory:

1. Deciding whether two tests measure the same trait or traits;
2. Analyzing the relative contributions of a pair of traits or abilities to test performance;
3. Detecting "functional" changes in items including those caused by security problems, mode of administration changes, and changes in familiarity with the concepts supporting the item in the population being tested;
4. Determining the adequacy of an item response function, i.e., a specific mathematical formula relating performance to ability;
5. Discovering the shape of the item response functions, including multidimensional item response functions;
6. Quantifying the magnitude and reliability of violations of the principal assumption of latent trait theory, "local independence";
7. Modeling item responses (such as omitting or changing answers) that fail to be locally independent.

In this paper theoretical results bearing on these problems will be outlined. A new theory will be presented, with the central problem the representation of traits, abilities or achievements, and their distributions.

1. Motivation

Three Practical Problems

To motivate the new theory, three important measurement problems will be used. First, however, two latent trait theory terms, "item response function" and "local independence," will be defined for the special situations considered in this section (to be redefined in later sections where more generality is needed).

The item response function (IRF; also called the item characteristic curve and conditional response function) is the (conditional) probability of sampling an examinee correctly answering the item from the subpopulation of all examinees at a particular ability level. Thus, if ability is unidimensional, the IRF for the i th item on a test is the real function P_i , where $P_i(t)$ = the probability of a correct response to item i from an examinee sampled from all those with ability = t . A pair of items, say the i th and j th, are said to be locally independent if they are independent in subpopulations having no variation in ability,

i.e., if for all ability levels t , e.g., $\text{Prob}\{\text{items } i \text{ and } j \text{ are both correct} \mid \text{ability} = t\}$ equals the product of the item response functions $P_i(t)P_j(t)$.

In the maintenance of testing programs that attempt to give more or less the same test, year after year, to a large number of people (e.g., military entrance and placement programs, college entrance exams, graduate and professional school admissions exams, high school and grade school aptitude and achievement tests, and interest measures such as job satisfaction scales used in industrial settings), three difficult measurement problems inevitably arise. These problems are functional item change, IRF adequacy, and local independence failure.

Functional item change. An item may function differently, i.e., have different psychometric properties, in two test administrations. For example, a vocabulary item requiring exposure to political terminology may seem relatively easy in a presidential election year. School curriculum changes, security problems, method of administration change, improper coaching, and item format change also may result in functional item change. The principal question is to determine to what extent, if at all, an item has functionally changed.

IRF adequacy. Many mathematical formulas have been proposed to represent IRFs. Psychological arguments have been used to challenge the correctness of each, usually over an ability range. For example, monotonic curves have been criticized for giving an incorrect representation over a low ability range because very low ability examinees may perform somewhat better than examinees just bright enough to be misled by item construction tricks. Curves that asymptote to 1.0 have been criticized because they contain no provision for the careless mistakes of very bright examinees answering items beneath their ability level. Of course, with the very large samples of examinees, virtually any guess or estimate of the population IRF can be rejected. The goal in adequacy problems is to determine whether a proposed curve is "adequate," i.e., close enough to the population IRF over an ability range to be acceptable in a specific application.

Local independence failure. Psychological reasoning or data analysis can sometimes lead to the suspicion that the local independence assumption has been seriously violated. For example, a pair of reading comprehension items referring to the same reading passage may be, to an unacceptable degree, measuring familiarity with the content of the passage. An example arising in an empirical item bias study is described later in this section. One application of the theory being developed is to determine the magnitude and reliability of suspected local independence failures.

Each of the three problems will be considered separately. Table 1 summarizes the discussion. It will be shown that all the problems can be reduced to a single question about two curves or functions of abilities: How close is a specified (either by a formula or table) function $L(\cdot)$ to an incompletely specified function $P(\cdot)$, the population IRF? Since abilities are only estimated and not observed, the question is difficult to answer.

IRF Adequacy

Consider the conceptually simplest problem type, IRF adequacy. A 3-parameter logistic function L has been estimated and offered as a representation of

Table 1
Summary of Formula Score Theory Analysis
of Three Basic Measurement Problems

Problem	Hypothesis	Population Parameter	Distribution of Test Statistic η
Functional Change	$L^* = P$	$\eta = \int_a^b (P-L)^2$	Quadratic function of normal variables; noncentral case.
Item Response Function Adequacy	$L = P$	$\eta = \int_a^b (P-L)^2$	Quadratic function of normal variables; central case.
Local Independence	$P_1 P_2 = P$ $P_1 = L_1$ $P_2 = L_2$	$\eta = \int_a^b (P-L_1 L_2)^2$	Quadratic function of normal variables; central case.

the true, i.e., population, IRF P . The psychometrician is concerned about the monotonicity of P and suspects that L fits P poorly over, say, the ability range $-3 \leq \theta \leq -2$. He/she wishes to determine how far apart P and L are over this range.

An intuitive and commonly used measure of the distance between two functions is the generalization of Euclidean distance given by the root mean square of function values. In this spirit, an attempt will be made to compute a point estimate and confidence interval for

$$\eta = \int_{-3}^{-2} [P(t) - L(t)]^2 dt . \quad [1]$$

The interval $[-3, -2]$ in the definition of η is arbitrary. The hypothesis being tested and the sample of examinees available for testing will generally suggest a different center and width of the "supporting" interval. Short intervals give more specific information about the difference between P and L . Very short intervals give estimates of η with a large sampling error.

The results of this section are made possible by an elementary equation which is valid at each point t where ability densities are continuous.

$$[P(t) - L(t)]f(t) = f^+(t)[1 - L(t)]\bar{P} + f^-(t)L(t)\bar{Q} \quad [2]$$

where

- P and L are as defined previously;
- f is the density for the ability θ in the general population of examinees;
- f^+ is the ability density in the subpopulation of examinees passing the target item;
- f^- is the conditional density in the failure subpopulation;
- and

$\bar{P} = 1 - \bar{Q}$ is the proportion of item passers.

This equation is important because it permits the evaluation of adequacy questions without estimating abilities. It is necessary to do so because for a test of fixed length, any estimate of ability has a substantial standard error for which a bound can be computed by routine methods. On the other hand, subject to technical qualifications treated at length below, an arbitrarily accurate estimate of the distribution of abilities can be obtained with a sufficiently large sample of examinees, and test administrations of over 1,000,000 examinees are no longer uncommon. In later sections, consistent estimates of ability densities are discussed.

In view of the very large sample sizes, f will be regarded as known. The effects of small errors in specifying f on the sampling distribution of the estimate of η has not been worked out at this time.

The quantity $\bar{P}$ on the right-hand side can either be computed as $\int L(t)f(t)dt$ when the hypothesis $P = L$ is being evaluated or be estimated as the sample proportion of persons who correctly answered the item.

In most applications only moderately large samples are available for estimating the conditional densities f^+ and f^- . Using Equation 2,

$$\eta = \int_a^b \{f^+(t)[1-L(t)]\bar{P} + f^-(t)L(t)\bar{Q}\}^2 W(t) dt \quad [3]$$

can be written, where the weight function $W(t)$ is $1/[f(t)]^2$. Upon substituting estimates $\hat{f}^+(\cdot)$ and $\hat{f}^-(\cdot)$ for $f^+(\cdot)$ and $f^-(\cdot)$, a statistic $\hat{\eta}$,

$$\hat{\eta} = \int_a^b \{\hat{f}^+(t)[1-L(t)]\bar{P} + \hat{f}^-(t)L(t)\bar{Q}\}^2 W(t) dt \quad [4]$$

is obtained that can be used to evaluate IRF adequacy. It will be seen that $\hat{\eta}$ has a tractable sampling distribution.

To motivate some theoretical developments on density representation and estimation in the next section, suppose the conditional densities could be represented in the form

$$f^+(\theta) = \sum_{j=1}^J \alpha_j^+ h_j(\theta) \quad [5]$$

and

$$f^-(\theta) = \sum_{j=1}^J \alpha_j^- h_j(\theta) \quad [6]$$

for known functions $h_1, h_2, \dots, h_J$ and constants α_j^+, α_j^- . Then, after the indicated integration in Equation 3 is carried out, η has the particularly simple form

$$\eta = \alpha Q \alpha^T \quad [7]$$

where α is the vector $\langle \alpha_1^+, \dots, \alpha_J^+, \alpha_1^-, \dots, \alpha_J^- \rangle$ and Q is a matrix of numbers that can be calculated prior to data collection. The entries in Q are obtained by substituting Equation 5 and 6 into Equation 3, expanding the product and numerically calculating the integral of the product of the specified functions. It is easily verified that Q is symmetric, positive, and semidefinite.

Such a representation has been derived. The functions h_j are derived from a priori considerations given in the next section. The number of them, J , turns out to be acceptably small, between 4 and 8, for the tests already analyzed.

Consistent, unbiased estimates for the vector of constants are described in the following sections. With them are obtained estimated densities

$$\hat{f}^+(\cdot) = \sum_{j=1}^J \hat{\alpha}_j^+ h_j(\cdot) \quad [8]$$

and

$$\hat{f}^-(\cdot) = \sum_{j=1}^J \hat{\alpha}_j^- h_j(\cdot) . \quad [9]$$

Here $\hat{\alpha}_j^+$ and $\hat{\alpha}_j^-$ are estimates of the corresponding constants.

The vector of estimates $\hat{\alpha}$ will be seen to be multivariate normal, at least asymptotically. The hypothesis $P = L$ permits calculation, prior to data collection, of the covariance matrix of the estimates and the derivation of the distribution of

$$\hat{\eta} = \hat{\alpha} Q \hat{\alpha}^T \quad [10]$$

Random variables of form $\hat{\alpha} Q \hat{\alpha}^T$, where $\hat{\alpha}$ is multivariate normal and Q positive, semidefinite, generalize the χ^2 family of random variables. In the "central case" the statistic has the same distribution as the sum of squares of several independent normal variables with zero mean and not necessarily equal variances. In the "noncentral case," the means may be unequal. The asymptotic normality and the hypothesis $P = L$ make the central case appropriate. A numerical algorithm has been developed by the author (with Bruce Williams) for computing the cdf $F(x) = \text{Prob}\{\hat{\eta} < x\}$ and determining the probability of observing an $\hat{\eta}$ equal to the sample value or larger under the hypothesis, $\eta = 0$, i.e., $P = L$. (For a review of alternative algorithms, see Johnson & Kotz, 1970, chap. 29. Technical details on the derivation and distribution of η are in Levine, 1981.)

Functional Item Change

The above approach can be used to attack functional item change questions. In the treatment of adequacy, the distribution of $\hat{\eta}$ was derived under the hy-

pothesis $P = L$. In studying change, the discrepancy between P and L is measured under the hypothesis that $P = L^*$ where L^* is some specified function other than L . Suppose, for example, that an IRF has been carefully measured using a very large sample and that years of successful experience with the item were consistent with the IRF used to represent it. The hypothesis $P = L^*$ has considerable support. However, after a security problem comes to light, a reestimation with a smaller sample gives a function $L \neq L^*$. Further, suppose that only low ability examinees are motivated to exploit the security problem. Under the hypothesis $P = L^*$, how large is the squared difference between P and L expected to be over the low ability range? By a generalization of the arguments described above, the distribution of $\hat{\eta}$ can be derived. It turns out to be a quadratic function of normal variables, noncentral case. Formulas for the variances and noncentrality parameters are in Levine (1981).

Local Independence

The method of this section suggests a way to quantify suspected departures from local independence. For example, in an item bias study in progress, a vocabulary item using the word "hurl" was found to be severely and reliably biased against sixth-grade girls and in favor of sixth-grade boys in two independent samples of 4,000 and 2,000 children. It seems likely that performance on another item also using a word favored by baseball writers would agree more with the hurl item score than that predicted by the local independence assumption of latent trait theory. To analyze the causes and consequences of bias, it would be valuable to have a method for measuring the magnitude and reliability of local independence violations over specified ability ranges.

To test for local independence, two suspect items may be (conjunctively) paired to form a complex item--an item that is scored correct if both component items are correct and incorrect otherwise. If the items are locally independent, then the item characteristic curve (ICC) of the complex item will be the product of ICCs of the component items. Evidence for a violation of local independence would be small

$$\eta_1 = \int_a^b [P_1(t) - L_1(t)]^2 dt \quad [11]$$

and

$$\eta_2 = \int_a^b [P_2(t) - L_2(t)]^2 dt \quad [12]$$

but large

$$\eta_{12} = \int_a^b [P_{1\&2}(t) - L_1(t)L_2(t)]^2 dt \quad [13]$$

In all the above examples, densities and conditional densities were represented as linear combinations of a finite set of known functions. It will be shown in later sections that every density is, in a sense soon to be made precise, equivalent to exactly one of these linear combinations. Every test will

be shown to have associated with it a unique "canonical space" or vector space of functions equivalent to densities. It is hoped that this discussion shows that a theory for density representation and estimation can be used to attack fundamental issues in psychological measurement.

2. Foundations: Canonical Space and Equivalent Ability Distributions

In the preceding section a relation was noted between several difficult but important substantive psychological issues and the more routine methodological problem of density estimation. The approach of the previous section required a representation of ability densities as finite linear combinations of known functions and multivariate normal estimates of the coefficients in the combinations.

This section informally reviews an a priori derivation of this representation; a more formal presentation of the derivation is outlined in Section 3. This approach to psychometric problems will be called "formula scoring" or "formula score theory" and abbreviated FS and FST.

The analysis is organized about three fundamental theoretical issues and methodological problems:

1. Ability distribution equivalence. Which, if any, pairs of fundamentally different ability distributions lead to exactly the same probability distributions on the item scores--the only observables in testing? What are necessary and sufficient conditions for two distributions to be equivalent (in the sense of making the same predictions)? What statement about the distribution are (in the technical foundations-of-measurement sense of the term) meaningful?
2. Ability distribution representation. Find a decomposition of an arbitrary ability density into two uniquely determined parts

$$f(\cdot) = f_0(\cdot) + f^*(\cdot) \quad [14]$$

such that two densities f_1 and f_2 are equivalent if and only if $f_1^* = f_2^*$.

Find a finite dimensional parameterization of the "identifiable part" f^* of the ability density f .

3. Ability distribution identification and estimation. Show that the "identifiable part" of the ability density is identifiable in the sense that a consistent estimate of $f^*(t)$ exists for each t . Construct an estimator.

The results of this section are derived from a version of latent trait theory that is more formally presented in Section 3.

The major random variables of the latent trait model are abilities θ and item scores $u_1, u_2, \dots, u_n$. Examinees are considered to be randomly sampled from an infinite population of examinees. The "points" in the probability space of the basic latent model are examinees. Each examinee has a specific ability and (nonrandom) vector of item responses. Abilities and item responses are non-trivial random variables only because examinees are sampled. Item responses are assumed to be "locally independent", i.e., independent in the subpopulations of examinees defined by conditioning upon ability. Although it may not be immediately obvious, this conceptualization of latent trait theory is compatible with the usual treatment of item responses as independent binomial random variables

with success probabilities that are functionally dependent on abilities, provided no item is ever administered two times to the same examinee.

To attack the problems of identifying, representing, and estimating ability distributions from a foundations-of-measurement point of view, the set of all statistics for a test is studied. Since only the item scores u_i are observed, and since examinees work independently of one another, the set of all statistics is simply the set of number-valued functions of the item score random variables. Moreover, this set can be shown to be a finite dimensional vector space.

An important tool for studying ability distributions in formula score theory is the canonical space of a test, formulated by referring to regression functions. The regression function of a statistic S is the real function

$$R_S(t) = E[S|\theta = t]. \quad [15]$$

The canonical space (CS) of a test is the vector space of all regression functions. It is easily shown to be finite dimensional. In fact, in many FST applications it has been possible to treat it as a vector space of low (less than 8) dimensionality. (See Section 4 for further discussion of CS dimensionality.)

Before proceeding, several assumptions commonly used in FST are listed. First, the functions

$$P_i = R_{u_i} \quad [16]$$

and

$$\begin{aligned} P_i(t) &= \text{probability that item } i \text{ is answered correctly given an} \\ &\quad \text{examinee with ability equal to } t \text{ has been sampled} \\ &= E(u_i|\theta = t) \end{aligned} \quad [17]$$

are assumed to be continuous. In addition, all abilities are assumed to lie in a closed bounded interval I . The assumption of continuous P_i is restrictive and may have to be dropped for some applications. The assumption of bounded abilities, on the other hand, results in no loss of generality because any latent trait model can be reformulated by routine methods as an isomorphic model with bounded abilities. These assumptions together imply that the canonical space consists of continuous functions on the interval I .

A major result of FST is that two densities are equivalent in the sense of problem 1 above if they have the same projection into the canonical space. Thus, if some J functions $h_1, h_2, \dots, h_J$ form a basis for the CS and if

$$\int_I h_j(t) f_1(t) dt = \int_I h_j(t) f_2(t) dt \quad j = 1, 2, \dots, J \quad [18]$$

then there is no objective way to choose between f_1 and f_2 .

By an "objective way to choose between f_1 and f_2 " is meant a method of using the observables (the item scores) to decide which of f_1 or f_2 is more nearly correct. This is impossible because it can be proven that every statistic has

the same probability distribution when f_1 is correct as when f_2 is correct.

This fact leads to a useful representation of densities. An arbitrary density $\underline{f}$ can be represented uniquely as

$$\begin{aligned} f(\cdot) &= f_0(\cdot) + \sum_{j=1}^J \alpha_j h_j(\cdot) \\ &= f_0(\cdot) + f^*(\cdot) \end{aligned} \quad [19]$$

where $\{h_j\}$ is a basis for the CS and f_0 is orthogonal to the regression function of every statistic, i.e.,

$$\int_I E(S|\theta = t) f_0(t) dt = 0 \quad [20]$$

for every statistic S . In this decomposition, f_0 is called the null part of $\underline{f}$, f^* the identifiable part of $\underline{f}$, and α_j the j th coordinate of $\underline{f}$.

The identifiable part of $\underline{f}$ is indeed identifiable because a sequence of estimators $\{\hat{f}_N(t)\}$ can be constructed that will (almost surely) converge to $f^*(t)$ as sample size N is increased. The convergence turns out to be uniform in t .

The null part of $\underline{f}$ is null in the sense that it is totally unrelated to data. There is no objective way to use the administered items to distinguish two densities with the same identifiable parts and different null parts. Such densities cannot and, for most purposes, need not be distinguished. Both densities lead to the same predictions in all applications. A proposition or scientific statement that is true if f_1 is the ability density and false if f_2 is the ability density is (in the technical foundations-of-measurement sense of the term) not meaningful. The proposition may be interesting, clearly stated and important, but there will be no way to tell if it is true or false from the observed responses to the test items.

The representation leads to a strategy for estimating densities. If the h_j (called coordinate functions) are orthonormal, then the coordinates α_j have a statistical interpretation,

$$\alpha_j = E[h_j(\theta)] , \quad [21]$$

i.e., α_j is the expected value of a function of the unobserved ability θ .

Since h_j is in CS, h_j is the regression function of some statistic, say X_j , and its regression function, the conditional expected value of X_j , will be equal to h_j :

$$E(X_j|\theta = t) = h_j(t) . \quad [22]$$

Since X_j is simply a function of item scores, X_j can be computed for each examinee in a large sample of, say, N examinees to obtain a sample mean $\bar{X}_j$. By the law of large numbers the estimate $\hat{f}_N(t)$

$$\hat{f}_N(t) = \sum_{j=1}^J \bar{X}_j h_j(t) \quad [23]$$

will converge (in probability) to the identifiable part of f as the sample size N becomes large.

This estimate is especially well behaved and easy to study because the examinees are independently sampled. In fact, for sufficiently large sample size N the vector of sample averages $N^{1/2} \langle \bar{X}_1, \bar{X}_2, \dots, \bar{X}_J \rangle$ will be nearly multivariate normal with covariance matrix that, at least in some applications, can be regarded as known.

3. An Outline of Some Basic Theory

For clarity and ease of future reference, Section 2 is outlined with the introduction of some additional assumptions and notation.

Basic Latent Trait Model and Notation (Simplest One-Dimensional Version)

- $\Omega = \{\omega\}$ = the probability space
 - = an infinite set of actual or conceivable examinees available for sampling and testing
- θ = ability random variable. $\theta(\omega)$ is unobserved.
- $f(\cdot)$ = the density for θ . Its support will always be assumed to be contained in an interval I . Except when noted, only continuous densities are considered.
- I = a closed interval containing all abilities:

$$\int_I f(t) dt = 1. \quad [24]$$

- n = number of test items.
- U = a random n -vector of item scores (the only observables)
 - = $\langle u_1, u_2, \dots, u_n \rangle$.
- u_1 = item score random variable. $u_1(\omega)$ is either zero or one.
- $P_1(\cdot)$ = item response function, or item characteristic function.
- $P_1(t) = E(u_1 | \theta = t) . \quad [25]$

These functions are assumed to be continuous.

Local independence assumption: For any n vector of zeros and ones

$$U^* = \langle u_1^*, u_2^*, \dots, u_n^* \rangle, \quad [26]$$

$$\text{Prob}\{U = U^* | \theta = t\} = \prod_{i=1}^n \{u_i^* P_i(t) + (1 - u_i^*) [1 - P_i(t)]\} \quad [27]$$

Statistic: A number-valued function of the item scores.

Basic Formula Score Terminology

Regression Function: The regression function of a statistic S is the conditional expectation

$$R_S(t) = E[S | \theta = t] . \quad [28]$$

Canonical Space (CS): The real vector space of all regression functions for a test. It is a finite dimensional subspace of the vector space of all continuous functions defined on I and can be shown to have dimension $\leq 2^n$.

J: The dimension of the canonical space (discussed in Section 4).

$(\cdot, \cdot)$: Notation for the inner product used on the space of continuous functions defined on I .

$$(g, h) = \int_I g(t)h(t)dt . \quad [29]$$

Note that

$$(h, f) = E[h(\theta)] . \quad [30]$$

Coordinate Functions: An orthonormal basis for the CS of a test, generally denoted $\{h_1, h_2, \dots, h_J\}$.

α_j : The projection of the ability density on the j th coordinate function, h_j . It is called the j th coordinate of f and has statistical interpretation

$$\alpha_j = E[h_j(\theta)] . \quad [31]$$

Ability Density Equivalence

$\chi_A(\cdot)$: The indicator function of the Set A

$$\chi_A(t) = 0 \text{ if } t \text{ is not in } A$$

$$= 1 \text{ if } t \text{ is in } A . \quad [32]$$

$P[., S, .]$: Notation for the probability distribution of the statistic S .

$P[A; S, f_1]$ is the probability that the statistic S is in the Set A when f_1 is the density for θ .

$$P[A; S, f_1] = \int E[\chi_A(S) | \theta = t] f_1(t) dt \quad [33]$$

Equivalent Densities: Two densities f_1, f_2 are equivalent for every statistic S

$$P[\cdot; S, f_1] = P[\cdot; S, f_2] \quad [34]$$

i.e., if the probability distribution of each statistic is the same when $f = f_1$ as when $f = f_2$.

Characterization of Equivalent Ability Densities: f_1 is equivalent to f_2 if and only if

$$(f_1, h_j) = (f_2, h_j) \text{ for } j = 1, 2, \dots, J \quad [35]$$

for any set of coordinate functions $\{h_j\}$.

Ability Density Decomposition

$g = g_0 + g^*$: Every density g on I can be expressed uniquely in the form:

$$g(\cdot) = g_0(\cdot) + g^*(\cdot) \quad [36]$$

where g^* is in the canonical space and for every statistic S

$$(R_S, g) = (R_S, g^*) \quad [37]$$

g^* : The identifiable part of a density g . The projection of the density into the CS.

$$g^*(t) = \sum_{j=1}^J \alpha_j h_j(t) \quad [38]$$

for coordinate functions $h_1, h_2, \dots, h_J$.

g_0 : The null part of a density g

$$g_0 = g - g^* \quad [39]$$

cannot and generally need not be estimated because $f_1^* = f_2^*$ implies $P[\cdot; S, f_1] = P[\cdot; S, f_2]$ for every statistic S .

Ability Density Representation

Densities and J vectors: The mapping

$$g \rightarrow \langle (g, h_1), (g, h_2), \dots, (g, h_J) \rangle \quad [40]$$

associates each density on I with a unique J vector.

Densities associated with the same J vector are equivalent.

Consistent, Unbiased Estimates of the Identifiable Part of the Ability Density

X_j : A statistic with regression function equal to h_j , that is,

$$R_{X_j}(\cdot) = h_j(\cdot). \quad [41]$$

There must be at least one because h_j is in the CS, and the CS consists of regression functions only. X_j must be bounded because it has all of its probability on a set of 2^n points.

$E(X_j) = \alpha_j$: Follows from

$$\begin{aligned} E(X_j) &= E[E(X_j|\theta)] \\ &= E[R_{X_j}(\theta)] \\ &= E[h_j(\theta)] \\ &= (h_j, f) \end{aligned} \quad [42]$$

$\bar{X}_{j;N}$: Sample mean of X_j from a sample of N examinees.

$\bar{X}_N = \langle \bar{X}_{1,N}, \bar{X}_{2,N}, \dots, \bar{X}_{J,N} \rangle$: Sample mean of N bounded independent, identically distributed random vectors. Converges to $\langle \alpha_1, \alpha_2, \dots, \alpha_J \rangle$. Asymptotically, $N^{1/2}\bar{X}_N$ is multivariate normal.

$$\hat{f}_N(t): \hat{f}_N(t) = \sum_{j=1}^N \bar{X}_{j,N} h_j(t) \quad [43]$$

a consistent, unbiased estimate of the identifiable part of the ability density. Asymptotically, $N^{1/2}\hat{f}_N(t)$ is normal.

Construction of Coordinate Functions h_j and Coordinate Estimators X_j : See Section 4.

4. Implementation of Ability Density Results

In this section the abstract results given in Sections 2 and 3 will be applied to estimate densities and ICCs. This section is included to show in a general way how the theory is used to analyze data. The discussion is organized about four technical questions that commonly arise in response to presentations of the theory:

1. How are the coordinate functions $h_1, h_2, \dots, h_J$ determined in actual applications?
2. How are the statistics X_j specified?
3. What is the dimension J and how is it determined?

4. Can the calculations be arranged in a way to avoid very long, numerically unstable calculations?

A set of statistics $\{S_k\}$ is called complete if any statistic S can be written as a linear combination of finitely many of them. For example, the elementary formula scores $\{v_k\}$ formed by considering all products of item scores are complete. These are the scores:

$$\begin{aligned} &1 \\ &u_1, u_2, \dots, u_n \\ &u_1 u_2, \dots, u_{n-1} u_n \\ &\vdots \\ &u_1 u_2 \dots u_n. \end{aligned} \quad [44]$$

The regression functions

$$R_{v_k}(t) = E[v_k | \theta = t] \quad [45]$$

are simply products of the ICCs P_i .

A set of coordinate functions can be constructed from any complete set of statistics as follows. First a function on $I \times I$ is specified by the formula

$$H(s, t) = \sum_k R_{S_k}(s) R_{S_k}(t). \quad [46]$$

"Is specified" in practical terms means that a computer subroutine is written that accepts pairs of numbers s, t as input and returns the number given on the right-hand side as output. In the special case where $\{S_k\} = \{v_k\}$, the "elementary" formula scores, it can be shown that H has the easily calculated form

$$H(s, t) = \prod_{i=1}^n [1 + P_i(s)P_i(t)]. \quad [47]$$

This special case will be discussed after the general case is treated.

Using standard methods $H(s, t)$ is decomposed into a finite sum of products of orthonormal functions. More specifically, a set of positive numbers $\lambda_1 \geq \lambda_2 \geq \dots, \lambda_J > 0$ and orthogonal functions h_j , $j = 1, 2, \dots, J$

$$(h_j, h_{j'}) = \begin{cases} 1 & \text{if } j = j' \\ 0 & \text{otherwise} \end{cases} \quad [48]$$

are computed such that

$$H(s, t) = \sum_{j=1}^J \lambda_j h_j(s) h_j(t) \quad [49]$$

for all s, t in I . Just as the eigenvalues of a positive semidefinite matrix are determined by the matrix, the λ_j 's and the "rank" J are determined by H . It can be shown that every function in the CS is a linear combination of the h_j no matter which complete set of statistics is used to construct H . In other words, the h_j are coordinate functions, and J is the dimension of the CS.

To construct an estimator of the coordinate

$$\alpha_j = E[h_j(\theta)] , \quad [50]$$

note that h_j is an eigenfunction of the "linear operator" defined by H . In symbols,

$$\begin{aligned} \lambda_j h_j(t) &= \int H(s, t) h_j(s) ds \\ &= \sum_k R_{S_k}(t) \int R_{S_k}(s) h_j(s) ds \\ &= \sum_k R_{S_k}(t) (R_{S_k}, h_j) . \end{aligned} \quad [51]$$

If

$$R_{S_k}(t) = E[S_k | \theta = t] \quad [52]$$

is replaced by S_k/λ_j in this formula, a statistic X_j is specified:

$$X_j = \lambda_j^{-1} \sum_k S_k (R_{S_k}, h_j) . \quad [53]$$

Since $E(S_k | \theta = t)$ is $R_{S_k}(t)$, it follows that

$$E(X_j | \theta = t) = h_j(t) \quad [54]$$

and

$$E(X_j) = E[h_j(\theta)] . \quad [55]$$

Thus, the sample mean $\bar{X}_j$ is a consistent, unbiased estimator of the coordinate α_j .

J , the dimensionality of the CS of the test, was calculated by analyzing any complete set of statistics $\{S_k\}$. J turned out to be the "rank" of the linear operator

$$h \rightarrow \varphi(h) = \int \sum_k R_{S_k}(\cdot) R_{S_k}(t) h(t) dt = \int H(\cdot, t) h(t) dt . \quad [56]$$

Although the eigenfunctions h_j and the eigenvalues λ_j depend on the choice of $\{S_k\}$, J does not. However, some applications may lead to particular $\{S_k\}$ and subsequently to a decision to treat the CS as if it had dimension $J' < J$.

The typical situation arises when the problem is considered of selecting $\hat{f}$ so as to minimize a quadratic index of goodness of fit, such as

$$Q(\hat{f}) = \sum_k [\bar{S}_k - E(S_k; \hat{f})]^2 . \quad [57]$$

Here $\bar{S}_k$ is the sample average value of S_k and

$$E(S_k; \hat{f}) = \int_I R_{S_k}(t) \hat{f}(t) dt \quad [58]$$

is the predicted value of $\bar{S}_k$. If

$$H(s, t) = \sum_{j=1}^J \lambda_j h_j(s) h_j(t) , \quad [59]$$

then Q can be written in the form

$$Q(\hat{f}) = \sum_{j=1}^J \lambda_j [\hat{\alpha}_j - \bar{X}_j]^2 + \text{terms that are independent of } \hat{f} \quad [60]$$

where

$$\hat{\alpha}_j = (\hat{f}, h_j) \quad [61]$$

$$\bar{X}_j = \lambda_j^{-1} \sum_k \bar{S}_k (R_{S_k}, h_j) . \quad [62]$$

From Equation 60 it can be seen that the size of λ_j measures the degree of improvement of fit to a set of statistics $\{S_k\}$ that can be obtained by including one more term in the representation of the identifiable part of the ability density

$$\sum_{j' < j} \alpha_{j', h_{j'}, (\cdot)} . \quad [63]$$

If λ_j is very small or if $\bar{X}_j$ has very large sampling error, then we do not attempt to estimate the coordinate and proceed as if the canonical space has lower dimensionality than J .

In applications J' has been selected by computing the λ_j and treating very small λ_j 's as zero. As a check on the adequacy of this procedure the differences between

$$J' - \sum_{j=1}^J (g, h_j) h_j(\cdot) \quad [64]$$

and $g(\cdot)$ are examined for various functions g . The functions g generally considered are the P_i , selected regression functions, and various guesses about $f(\cdot)$. The two functions will agree exactly if g is in the CS and J' is the dimensionality of the CS.

One formula scoring technique has proven more accurate than all of the other techniques that have been tried. The complete formula score method (CFSM) uses the elementary formula scores $\{v_k\}$ as its complete set of statistics and begins with the identity

$$\begin{aligned} H(s, t) &= \sum_k R_{v_k}(s) R_{v_k}(t) \\ &= \prod_{i=1}^n [1 + P_i(s) P_i(t)] . \end{aligned} \quad [65]$$

(The identity is verified by induction or by expanding the product.) The sum has 2^n terms, but the product has only n terms and thus can be calculated with many fewer operations.

The X_j can also be calculated with a variant of this identity. Each sampled examinee's data is transformed to define a random continuous function $V(t)$, which is called the V-transform of the examinee's data. This function

$$V(t) = \prod_{i=1}^n [1 + u_i P_i(t)] \quad [66]$$

is easily calculated and is identically equal to

$$\sum_{k=1}^{2^n} v_k R_{v_k}(t) . \quad [67]$$

Therefore, $\int_1 V(t) h_j(t) dt$ equals

$$\sum_k v_k (R_{v_k}, h_j) = \lambda_j X_j . \quad [68]$$

This reduces the calculation of X_j from 2^n operations to one numerical integration. In fact, to calculate the sample mean $\bar{X}_j$, only one integration need be

done. This is true because the sample average of the integrated $V(t)$ is the integral of the sample average

$$\text{Ave}\{\int V(t)h_j(t)dt\} = \int [\text{Ave } V(t)]h_j(t)dt . \quad [69]$$

Thus, in applications, $V(t)$ is computed on a grid of t values for each examinee, accumulated over examinees and numerically integrated to obtain the J sample means $\bar{X}_1, \bar{X}_2, \dots, \bar{X}_J$. This procedure can be adapted to compute sample covariances of the X_j by numerical integration.

5. Discovering the Shapes of Item Response Functions

Sometimes, but not always, the shape of ICCs can be rationally deduced and parameterized. For example, the S-shape of the 3-parameter logistic curve on an unbounded ability continuum follows from

1. Monotonicity. More able examinees are more likely to answer correctly.
2. Asymptotes. Although the probability of a correct response can be made arbitrarily close to 1.0 by sampling from very high ability subpopulations, a substantial proportion of each low ability subpopulation will select the correct option of a well-constructed multiple-choice item.
3. Simplicity/Parsimony. The item response curve has no more points of inflection than the one implied by the monotonicity and asymptotes conditions above and by smoothness conditions.
4. Symmetry. The graph of the ICC is symmetric for high and low ability levels in the sense that a length-preserving transformation $(x,y) \rightarrow (2x_0-x, 2y_0-y)$ about a point of inflection on the graph (x_0, y_0) carries the curve into itself.

Fitting 3-parameter logistic functions is sensible when these conditions are met because every curve satisfying these conditions will be close to at least one 3-parameter logistic function.

Sometimes these assumptions are implausible or clearly false, and a method is needed to discover and parameterize shape. As a one-dimensional example, it was demonstrated (Levine & Drasgow, in press) with a very large sample of aptitude test examinees that the conditional response function for the response of choosing a particular (incorrect) option on a multiple-choice test for several items was clearly nonmonotonic. In multidimensional measurement the shape of the item characteristic surface is a matter of considerable psychological importance because it represents a statement about how several abilities interact to simultaneously determine response probability. In the next section, the methods discussed here are used to develop a procedure for determining the shape of an item characteristic surface.

Another application of the method described in this section is the study of item responses, such as omitting, that cannot reasonably be expected to satisfy local independence. FST permits the construction of a consistent estimate of an "omitting characteristic curve" $P[\text{item } i \text{ is omitted} | \theta = t]$, without assuming local independence for omitting responses.

The basic issues addressed in this segment of the research are the following:

1. What is the shape of a new item or item type's ICC?
2. What information about ability is contained in incorrect answers and item omitting?
3. Modeling item responses that may not be locally independent, such as omitting and the following examples drawn from computer-administered tests: (a) attempting to change an answer, (b) requesting a display of a previously presented item or part of an item, (c) responding in a time clearly too short to read the item, (d) responding with potentially damaging force to the terminal.

Equations Relating FST to ICC Shape

Our results in this area depend on the following elementary relation which is used to reduce IRF estimation to ability density estimation:

$$\text{Prob}\{\text{correct response} | \text{ability is in Set A}\} = \frac{\text{Prob}\{\text{ability is in A} | \text{correct response}\}}{\text{Prob}\{\text{ability is in A}\}} \times \text{Prob}\{\text{correct response}\} \quad [70]$$

Thus, the conditional response probability is proportional to the ratio of the ability distribution in the subpopulation of examinees correctly answering the item to the unconditional distribution. (The constant of proportionality, $\text{Prob}\{\text{correct response}\} = \bar{P}$, has an obvious consistent unbiased estimate, the sample proportion correct). If regularity assumptions are made, then

$$P(t) = \frac{f^+(t)}{f(t)} \bar{P} \quad [71]$$

where f^+ is the ability density in the subpopulation of item passers and f is the ability density. This equation can be written in the form

$$P(t) = \frac{1}{1 + \frac{\bar{Q} f^-(t)}{\bar{P} f^+(t)}} \quad [72]$$

where $\bar{Q} = 1 - \bar{P}$ and f^- is the failure density.

This formula is especially useful for developing a distribution theory for the estimates because the failure and passing subpopulations are disjoint. Thus, the distribution of an estimator of $f^+(t)/f^-(t)$ can be developed by studying the ratio of statistically independent random variables.

Several variations of this approach to ICC estimation have been tried with generally satisfactory but occasionally poor results. Systematic comparisons of the variations will be made after more exploratory work. Some current and projected refinements are delineated below.

1. If ability is uniformly distributed or if $f(\cdot)$ is constant on a range of abilities of interest, then the ICC is proportional to f^+ , and ICC estimation is simplified. Furthermore, sampling fluctuation is unlikely to give a zero or negative density estimate. An initial approximation can be used to transform ability so that $f(\cdot)$ is approximately constant.
2. The sampling distribution of coordinate estimates depends on the choice of $\{S_k\}$. It has been observed that if h_1 , the coordinate function with the largest eigenvalue is close to $f(\cdot)$, then very good estimates of $f(\cdot)$ are obtained. Currently, an attempt is being made to capitalize on this effect by carefully choosing $\{S_k\}$ and controlling the eigenfunction shapes.
3. The current density estimates are least squares in the sense that they minimize the residual sum of squares Q (Section 4). The results on density equivalence permit the expression of the likelihood function as a function of finitely many parameters, the coordinates a_j . In principle, this expression could be maximized and maximum likelihood density estimates could be computed.

The basic equation used to relate ICCs to density ratios is essentially the definition of conditional probability. Local independence plays no role in its derivation. In fact, the binary item response on the focal item is being used merely to divide the sample into those who answer the item correctly and those who answer it incorrectly. The equation would remain valid if any binary score were used to dichotomize the sample and population. Therefore, the same estimation techniques used to estimate ICCs could be used to study the relation of ability to complex item responses that failed to satisfy the local independence assumption. These include item skipping and very fast responding.

6. Multidimensional Formula Scoring for Homogeneous Subtests

There is an interesting and important multidimensional measurement problem that can be implemented with currently available unidimensional software. The problem is to discover how several abilities jointly determine response probability on new item types. This can be done when certain psychological assumptions are valid. In particular, it is necessary that a variety of item types are available, that all items depend on the same small number of abilities, and that items can be grouped into "homogeneous" subtests.

For concreteness consider three item types: synonyms, antonyms, and analogies. Suppose that all three depend only on a pair of abilities θ_1, θ_2 in the sense that for any ability levels s, t and any r items, the item scores $u_{11}, u_{12}, \dots, u_{1r}$ satisfy

$$E\left[\prod_{j=1}^r u_{1j} \mid \theta_1 = s \ \& \ \theta_2 = t\right] = \prod_{j=1}^r E[u_{1j} \mid \theta_1 = s \ \& \ \theta_2 = t] . \quad [73]$$

In other words, the items are independent in subpopulations formed by conditioning on both abilities.

Homogeneous subtests, defined more formally below, are unidimensional sub-

tests of a multidimensional test. For example, both synonym and antonym items may require both language fluency, θ_1 , and an ability to recognize and generalize abstract relations, θ_2 ; but the antonym items can be written in such a way as to demand a relatively large amount of the second ability. Thus, synonym items may satisfy local independence with respect to some linear or nonlinear function of θ_1 and θ_2 , say $\theta_1 + \theta_2$, and antonym items with respect to, say $\theta_1 + 2\theta_2$. Subtests consisting of one item only will appear unidimensional, but the total test will not.

Note that the assumption that item types from homogeneous subtests is more general (and more believable) than the assumption that different item types measure different traits. This should be obvious after the discussion of "ad hoc coordinates."

If certain plausible assumptions (described below) are correct, then unidimensional parameter estimation programs can be used with a test consisting only of synonym items and a test consisting only of antonym items. FST can then be used to represent and to estimate bivariate analogy IRFs.

Homogeneous Subtests and Ad Hoc Coordinates

Latent trait theory provides a way to precisely state what is meant by a homogeneous subtest and items requiring different amounts of unobserved abilities. As before, the population of examinees is denoted by a point set $\Omega = \{\omega\}$. In this situation, abilities map examinees into two vectors rather than numbers: $\theta(\omega) = \langle \theta_1(\omega), \theta_2(\omega) \rangle$. Because examinees are randomly sampled, θ is a random vector.

To quantify the notion of homogeneous subtests, a pair of number-valued functions ϕ, ψ are considered. The first subtest is homogeneous in the sense that it satisfies a local independence assumption relative to $\phi(\theta)$. In symbols, for items $i_1, i_2, \dots, i_r$ on the first subtest

$$\text{Prob}\{u_{i_1} = 1, u_{i_2} = 1, \dots \text{ and } u_{i_r} = 1 | \phi(\theta) = t\} = \prod_{j=1}^r P_{i_j}(t) \quad [74]$$

where

$$P_{i_j}(t) = \text{Prob}\{u_{i_j} = 1 | \phi(\theta) = t\} . \quad [75]$$

A similar equation expresses the assumption that the second subtest is homogeneous with respect to $\psi(\theta)$: The item scores for items in the second subtest are independent in the subpopulation of examinees with the property $\psi(\theta) = s$ for each constant s .

Note that these conditions permit using available parameter estimation programs to calibrate each subtest separately. Somewhat paradoxically, each item is essentially multidimensional, but each subtest satisfies the axioms of one-

dimensional latent trait theory. FST provides a method for integrating the subtests and modeling the analogy items that also depend on fluency and abstraction, but to an unknown extent.

The functions ϕ and ψ are called "ad hoc coordinates." If for each $\langle s, t \rangle$ there is at most one vector $\langle x, y \rangle$ satisfying $s = \phi(x, y)$ and $t = \psi(x, y)$, and certain regularity conditions are met, then ϕ and ψ can be treated as curvilinear coordinates for the set of (bivariate) abilities.

The FST analysis to be presented only gives a representation of item response function in terms of ad hoc coordinates

$$P_1(s, t) = \text{Prob}\{u_1 = 1 | \phi(\theta) = s \text{ \& } \psi(\theta) = t\} \quad [76]$$

For many modeling problems this is adequate. Admittedly, the variables θ_1 (fluency) and θ_2 (abstraction) are considerably more interesting than ϕ and ψ . Conjoint measurement or uniform systems analysis (Levine, 1970) may permit the analysis of the relation between ϕ, ψ and θ_1, θ_2 . However, the current concern is with the psychometric problem of representing new items in the ad hoc system.

Formula Score Approach to Representing New Item Types

Consider a 21-item test consisting of 10 synonym items followed by 10 antonym items and one analogy item. The task is to compute

$$\text{Prob}\{u_{21} = 1 | \phi(\theta) = s \text{ \& } \psi(\theta) = t\} = P(s, t) . \quad [77]$$

First, the homogeneous subtests will be analyzed separately. The most direct approach would be to first calibrate the synonym items by embedding them in a large conventional administration of many items of the same type. The analysis would yield ϕ ICCs

$$\text{Prob}\{u_1 = 1 | \phi(\theta_1, \theta_2) = t\} = P_1(t) \quad i = 1, 2, \dots, 10 . \quad [78]$$

Similarly, a separate analysis of the antonym items yields ψ ICCs

$$\text{Prob}\{u_1 = 1 | \psi(\theta_1, \theta_2) = t\} = P_1(t) \quad i = 11, 12, \dots, 20 . \quad [79]$$

To discover the shape of an analogy item's ICC, a 21-item test would be administered to a sample of examinees. By an obvious generalization in Section 5 the analogy item ICC

$$P(s, t) = \text{Prob}\{u_{21} = 1 | \phi(\theta_1, \theta_2) = s \text{ and } \psi(\theta_1, \theta_2) = t\} \quad [80]$$

can be represented as a ratio of densities.

$$P(s, t) = \frac{f^+(s, t)}{f(s, t)} \bar{P} \quad [81]$$

where f^+ is the (bivariate) density in the subpopulation of examinees who correctly answered Item 21 and $\underline{f}$ is the unconditional density.

Before continuing this analysis, the CS for the 20-item test is described. It turns out that the assumptions imply that the canonical space of the 20-item test is simply the "tensor product" of the CS for the first subtest and the second subtest.

At this point it seems advisable to restate some definitions. The CS for the first subtest is the set of one-dimensional regression functions

$$E[S|\phi(\theta_1, \theta_2) = t] = R_S(t) \quad [82]$$

where S is a statistic whose value depends on the first 10 item scores only. The CS for the antonym item is the set of regression functions $E[S|\psi(\theta_1, \theta_2) = t]$ for statistics S that are functions of the second 10 scores only.

The CS for the first 20-item subtest will be the set of regression functions

$$R_S(s, t) = E[S|\phi(\theta_1, \theta_2)] = s \quad [83]$$

$$\psi(\theta_1, \theta_2) = t \quad [84]$$

for the statistics S of the first 20 scores. The psychometric assumptions imply that any $R(s, t)$ in the 20-item CS can be written as a finite sum of functions of the form $h(s)h'(t)$ for h in the first subtest CS and h' in the second subtest CS. In fact if $\{h_1, h_2, \dots, h_J\}$ is a basis for the first CS and $\{h'_1, h'_2, \dots, h'_{J'}\}$ for the second CS, then the $J \times J'$ functions

$$h_j(s)h'_{j'}(t) = h_{jj'}(s, t) \quad [85]$$

where $j = 1, 2, \dots, J$ and $j' = 1, 2, \dots, J'$ will be a basis for the two-dimensional CS.

Current one-dimensional FS programs can be used to estimate the bivariate densities f^+ and $\underline{f}$. The estimate will have form

$$f^+(s, t) = \sum_{j=1}^J \sum_{j'=1}^{J'} \bar{X}_{jj'} h_{jj'}(s, t) \quad [86]$$

where the sample mean $\bar{X}_{jj'}$ is a consistent, unbiased estimator of $E[h_{jj'}(\theta)]$. If f^+ and $\underline{f}$ are in the 20-item CS, then (by sample splitting to estimate f^+ and $\underline{f}$ separately) a consistent estimate of $P(s, t)$ is easily specified, and the shape of the item response surface can be "discovered."

The success of this approach requires f^+ and $\underline{f}$ to be in or near the 20-item two-dimensional CS. This assumption seems plausible when the variety of shapes that can be constructed as linear combinations of the $J \times J'$ coordinate functions are considered.

6. Are Two Tests Measuring the Same Trait(s)?

Suppose a major revision is made of a complex, not necessarily unidimensional test. Does the new test measure the same traits? Suppose the format or mode of administration is changed. Does the test still measure the same traits? Suppose a translation of the test into another language is attempted and that it is unlikely that every original test item is psychometrically equivalent to its translation. Can the translated test nonetheless measure the same traits as the original? These questions lead to asking the foundations-of-measurement question,

What necessary and/or sufficient conditions must item scores obey before it can be concluded that two nonparallel tests are measuring the same trait(s)?

It is hoped that theoretical work on this problem will lead to a statistical test that can be used in applications. This section relates FST to the problem and reports some current work.

Two nonparallel tests are administered to the same population. (The tests can be considered as subtests of one test.) Item response curves are fitted to each test separately. In other words, except possibly to compute an equating transformation, only Test 1 scores are used to estimate the IRF of a Test 1 item. Can the variable in the first set of IRFs be given the same interpretation as the variable in the second set of IRFs?

The mathematical kernel of this problem seems to be this. A probability measure is given for a set Ω along with two sets of zero-one random variables: $u_1, u_2, \dots, u_n$ and $u_1', u_2', \dots, u_{n'}'$. Two sets of real functions, $P_1, P_2, \dots, P_n$ and $P_1', P_2', \dots, P_{n'}'$, are also given. What conditions must be assumed in order for it to be possible to construct one more random variable θ such that the P 's are IRFs for the u 's relative to θ and all $n + n'$ u 's are locally independent relative to θ .

A strong necessary condition on the given scores and functions can be formulated with FS notation and concepts. Let $\{v_k\}$ denote the elementary scores in the first set of u 's (Section 5). Let $\{v_{k'}'\}$ denote the elementary scores in the second set of u 's. Let R_{v_k} and $R_{v_{k'}'}$ be the corresponding products of P 's. Thus, if

$$v_k = u_{i_1} u_{i_2} \dots u_{i_r} \quad [87]$$

then

$$R_{v_k}(\cdot) = P_{i_1}(\cdot) P_{i_2}(\cdot) \dots P_{i_r}(\cdot). \quad [88]$$

Let $\{v_{k,k'}'\}$ denote the elementary formula scores for the $n + n'$ item test where $v_{k,k'}'$ is $v_k v_{k'}'$. Let $R_{k,k'}'$ denote the corresponding product of functions for $v_{k,k'}'$, that is,

$$R_{k,k'}' = R_{v_k} R_{v_{k'}'} \quad [89]$$

Let $\{h_j\}_{j=1}^J$ be an orthonormal basis for the linear span of the $\{R_{k,k^-}\}$. For example, the h_j could be obtained by analyzing the function $H(s,t)$ defined by

$$\prod_{i=1}^n [1 + P_i(s)P_i(t)] = \prod_{i=1}^{n'} [1 + P_i'(s)P_i'(t)] \quad [90]$$

as in Section 4. Finally, let X_j be the random variable

$$X_j = \sum_k \sum_{k^-} v_{k,k^-} (R_{k,k^-}, h_j) , \quad [91]$$

which can be more conveniently computed as described in Section 4.

It is easy to show the following condition on expected values of scores

$$E(v_k v_{k^-}) = \sum_{j=1}^J E(X_j)(h_j, R_{k,k^-}) \quad [92]$$

is necessary for all elementary formula scores $v_k v_{k^-}$.

The condition also appears to be sufficient, although a proof is not available at this time. In any event, additional work is needed to determine when one random variable suffices for Ω and specified subpopulations of Ω .

ACKNOWLEDGMENTS

This research was supported by Office of Naval Research Contract N00014-79C-0752, NR 150-445.

REFERENCES

- Johnson, N.L., & Kotz, S. Continuous univariate distributions (Vol.2). Boston: Houghton-Mifflin, 1970.
- Levine, M. V. Transformations that render curves parallel. Journal of Mathematical Psychology, 1970, 7, 410-443.
- Levine, M. V. Tests of local independence and item fit (Working Paper 81-1). Champaign-Urbana: University of Illinois, Department of Educational Psychology, 1981.
- Levine, M. V., & Drasgow, F. The relation between option choice and estimated ability. Educational and Psychological Measurement, in press.